Visual Tracking of Hand Posture in a Robot Control Application

Fabienne Lathuilière Jean-Yves Hervé
Groupe de Recherche en Perception et Robotique
École Polytechnique, Montréal, QC H3C 3A7
tel (514) 340-4711-5783, fax (514) 340-5802,
email: [fabienne,jyh]@ai.polymtl.ca

Abstract

The properties of versatility and dexterity of the human hand have raised a growing interest both in Human Computer Interaction applications and in gesture recognition. Emphasis is laid on ease of detection and reconstruction of the hand posture as well as on real-time computation. This paper proposes a visual hand tracking system and a hand posture estimation method where the position and the orientation are recovered from interest cues on the hand. A kinematic model of the hand and a reconstruction method to recover hand posture are provided. A skeletal model of the hand is built to model the kinematic properties of the hand. A single camera provides the frame sequence of the mobile hand while color segmentation is used to detect salient features on the glove adorned hand. After the pose of the wrist is computed, the value of the finger joint angles are obtained by inverse kinematics. This method enables the successful recovery of the hand posture, opening the door to applications such as 3D input devices or powerful 6-dof control tools. Human hand skill can thus be exploited to control a robot gripper in a master-slave system taking advantage of the information provided by the human guide.

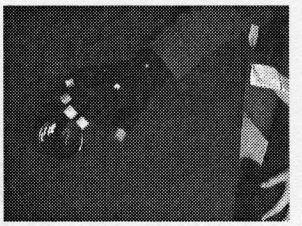
1 Introduction

Recent developments in Human Computer Interaction (HCI) have led to different approaches as regards the use of the hand as an input device. Sensor-equipped devices such as input gloves or magnetic sensors hinder the user's natural movements and furthermore lack precision. In order to avoid such cumbersome and expensive devices, a lot of work has been dedicated to visual interfaces with one or more cameras tracking the hand. A recent survey about HCI [1] highlights the requirements of an efficient visual interface device. In particular, the ease of use, the absence of tedious calibration steps, and real-time reaction of the system are of much importance, hence the current efforts to create efficient vision-based interface applications. Two main categories stand up in vision-based applications. The first class of work places the emphasis on gesture recognition, where only a finite amount of hand gestures are recognized. For instance, the system presented in [2] recognizes several American Sign Language (ASL) gestures by tracking fingertip movement. Each gesture is defined by the vector list of the fingertip motion, and recognition is achieved by look-up table list matching. [3] uses a view-based representation of the hand to learn and recognize dynamic hand gestures by statistical matching. In both cases, the palm remains in a constant location and no 3D tracking is addressed. The second approach emphasizes gesture reconstruction, possibly through the definition of some specific mapping between hand posture and actions on the device. In this case, a 3D model of the hand is needed and hand posture is generally recovered by a fitting process. Among these applications, [4] uses a Kalman filter to determine 3D hand position, and [5] proposes a tracking system of the index finger by 2D image fitting for a virtual gun interface, but both systems require some part of the operator to remain still. [6] successfully recognizes the posture of an unadorned hand by model state estimation and residual vector minimization, but needs two cameras to successfully position in space characteristic lines and points on the fingers. Finally, the system described in [7], based on color markers to designate the fingertips, leads to a precise reconstruction of the hand posture but spends much time in iterative procedures. The approach chosen in this article aims at providing enough information on an adorned hand to recognize its posture in space with a single camera and little com-

The support of NSERC (RGPIN 171210-97) for this work is gratefully acknowledged.

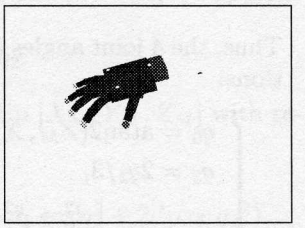
putational burden. Indeed, the increasing amount of computers with one camera mounted on top of the monitor urges the development of easy-to-use visual interfaces. The hand is covered with a black glove adorned with color tips, letting the user do natural movements. A mathematical model of the hand is defined as well as its kinematic properties to develop a reconstruction method of the hand posture. Experiments have proved the efficiency of the detection of the features and the good recovery of the hand posture, enabling to animate a graphical replica of the hand while the hand is in motion. The procedure is shown in Figure 1.

(a) experimental setup



(b) illuminated hand

(c) gloved hand in operating conditions



(d) reconstructed hand

Figure 1: The hand detection system and the reconstruction procedure

This paper will first address the kinematic model of the hand, the steps in detecting and reconstructing the hand and the animation of the virtual model. Finally, the visual interface to control the robot gripper is considered.

2 The kinematic model of the hand

The hand model is designed in order to remain simple enough to respect human capabilities and span of motion. Some degrees of agility are reduced but need not restrict the inverse pose problem. The hand is modeled by a 26 degree of freedom skeleton, whose location is given by the the wrist's middle point and whose orientation is given by that of the palm. The fingers are enumerated from I to V from the thumb to the little finger. Each finger has 4 degrees of freedom, namely one in abduction/adduction and 3 in flexion/extension. Inspired from [8], the segments of articulation of each finger are concurrent at the wrist's middle point, C, as shown in Figure 2.

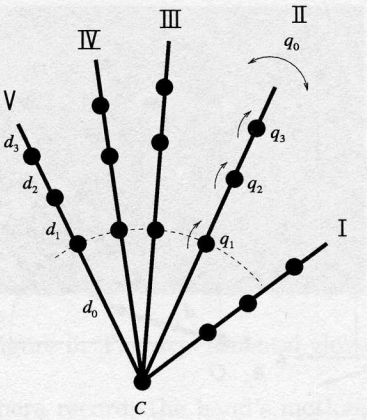


Figure 2: Skeletal model of the hand

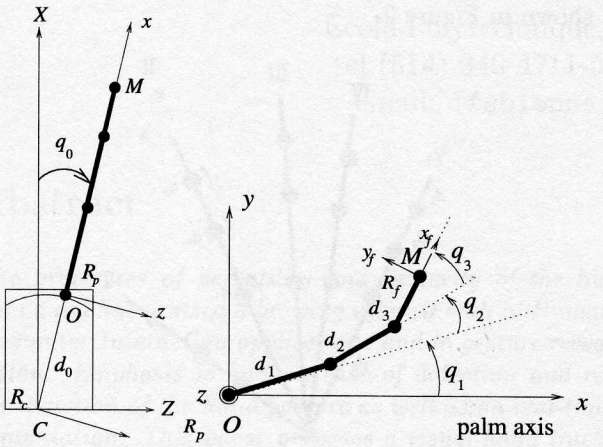
The abduction angle characterizes the angle of the finger in the palm's plane, whereas the flexion angle corresponds to the folding of the finger in the perpendicular plane of the palm. Each finger but the thumb is assumed to be a planar manipulator. All the corresponding planes of motion of the fingers intersect at the middle point of the wrist.

2.1 Case of fingers II to V

We will be using three reference frames for representing the kinematics of the fingers:

- $R_c = [C, X, Y, Z]$ is the hand's reference frame, whose origin is the wrist's middle point and whose axes are defined by the palm's orientation, as shown in Figure 3(a).
- $R_p = [O, x, y, z]$ is the finger plane's reference frame whose origin is at the first flexion/extension joint and whose x-axis is the direction of the finger in the palm's plane, as shown in Figure 3(a).
- $R_f = [M, x_f, y_f, z_f]$ is the finger's final frame, where M is the finger tip and axes x_f and y_f lie in the plane of motion, as shown in Figure 3(b). The third axis, z_f, is obviously parallel to the z-axis of the finger plane's frame.

Each finger is modeled as a 4-dof planar manipulator. The abduction angle, q_0, is defined as the angle between the finger's plane of motion and the axis of reference X. The flexion angles, q_1, q_2, and q_3 respectively represent the metacarpophalangeal, proximal interphalangeal, and distal interphalangeal joint angles, as shown in Figure 3.



(a) Top view (b) View in the plane of motion

Figure 3: Kinematic model of the fingers II to V

Direct kinematics of the finger leads to the following homogeneous transformation matrix defining the position and orientation of the finger's final frame R_f with respect to the hand's reference frame R_c, where $T[a,d]$ stands for a translation of d along axis a and $R[a,\theta]$ stands for a rotation of angle θ around axis a:

$$[M]_{R_c} = R[Y,q_0]\ T[x,d_0]\ R[z,q_1]\ T[x,d_1]\ R[z,q_2]\ T[x,d_2]\ R[z,q_3]\ T[x,d_3]\ [M]_{R_f}.$$

Following the simplification proposed by [9], we linearize the relationship between the angles q_2 and q_3 such that

$$q_3 = \frac{2}{3}\ q_2.$$

This simplification leads to a direct solution for the inverse kinematics problem of the finger posture. Assuming $[M]_{R_p}$ given by $[X_M, Y_M, Z_M]$, the flexion angles can be computed according to the representation given in Figure 4. The triangle relationship leads to:

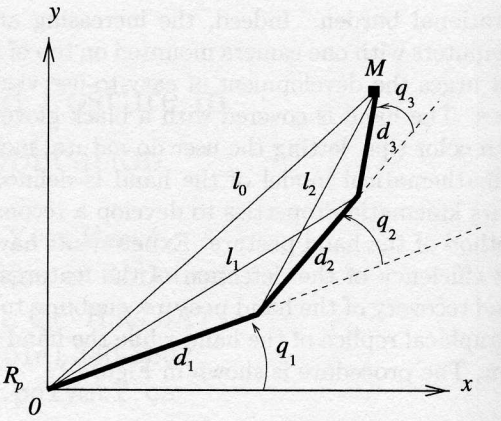


Figure 4: Inverse kinematics of the finger

$$\begin{cases} l_0^2 = X_M^2 + Y_M^2, \\ \alpha = \text{atan2}(X_M, Y_M) - (\frac{\pi}{2} - q_1 - q_2 - q_3), \\ l_1^2 = d_1^2 + d_2^2 + 2d_1 d_2 \cos q_2 \\ \quad = l_0^2 + d_3^2 - 2l_0 d_3 \cos\alpha, \\ l_2^2 = d_2^2 + d_3^2 + 2d_2 d_3 \cos q_3 \\ \quad = l_0^2 + d_1^2 - 2l_0 d_1 \cos[\text{atan2}(Y_M, X_M) - q_1]. \end{cases}$$

Thus, the 4 joint angles are given by the following equations:

$$\begin{cases} q_0 = \text{atan2}(Z_M, X_M), \\ q_3 = 2q_2/3, \\ \beta = \text{acos}\left[\dfrac{l_0^2 + d_1^2 - d_2^2 - d_3^2 - 2d_2 d_3 \cos q_3}{2l_0 d_1}\right], \\ d_1^2 + d_2^2 - l_0^2 - d_3^2 + 2d_1 d_2 \cos q_2 \\ \qquad +2l_0 d_3 \cos 5q_2/3 - \beta = 0, \\ q_1 = \text{atan2}(Y_M, X_M) - \beta, \end{cases}$$

where q_2 is obtained by numerical computation.

2.2 Case of the thumb

The thumb has one additive degree of freedom at the abduction/adduction proximal interphalangeal joint. In order not to add tedious computation to the inverse kinematics problem, we chose to keep 4 degrees of freedom to describe the motion capabilities of the thumb. The resulting motion restriction does not alter the thumb's gripping capabilities, as far as robotic applications are concerned. The motion of the thumb is defined by one rotation around the z-axis followed by 2 rotations around the y-axis, in relative frame description. The thumb tip thus points to the little finger knuckle base (see Figure 5).

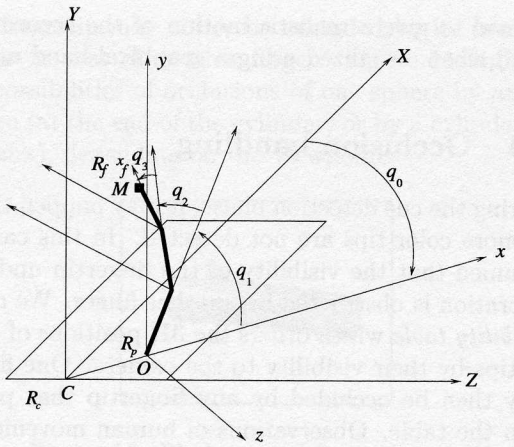


Figure 5: Kinematic model of the thumb

The transformation matrix is then given by:

$$[M]_{R_c} = R[Y, q_0]\ T[x, d_0]\ R[z, q_1]\ T[x, d_1]$$
$$R[y, q_2]\ T[x, d_2]\ R[y, q_3]\ T[x, d_3]\ [M]_{R_f}.$$

We link the last two angles according to the following empirical equation, which appears to be a good approximation of reality:

$$q_3 = q_2.$$

The position of the thumb's tip $[X_M, Y_M, Z_M]$ with respect to R_p can be written:

$$\begin{cases} X_M &= \cos q_1 (d_1 - d_3 + d_2 \cos q_2 + 2d_3 \cos {q_2}^2), \\ Y_M &= \sin q_1 (d_1 - d_3 + d_2 \cos q_2 + 2d_3 \cos {q_2}^2), \\ Z_M &= -\sin q_2 (d_2 + 2d_3 \cos q_2). \end{cases}$$

In the same frame, the inverse kinematics expression comes forward:

$$\begin{cases} q_1 = \text{atan2}(Y_M, X_M), \\ q_2 = \text{acos} \dfrac{-d_2 + \sqrt{d_2^2 - 4(2d_3(d_1 - d_3 - X_M/\cos q_1))}}{4d_3}, \\ q_3 = q_2. \end{cases}$$

3 Hand detection and reconstruction

3.1 Cue detection

A minimum number of color cues need to be added on the hand to perform accurate posture reconstruction. Three points on the palm are first necessary to recover both its position and orientation. Second, since the final position of one fingertip is sufficient to compute the knuckle angles provided the abduction angle is known, each fingertip is marked with a unique color. The user thus wears a dark glove covered with one colored label on each fingertip and three on the upper-palm. Among the hue-equidistant colors in the Hue Saturation Value system, bright colors appeared to be more suited to color segmentation under non uniform lighting. As a result, white points were chosen for the palm then one red, orange, cyan, green and yellow point for the different fingers, as shown in Figure 6.

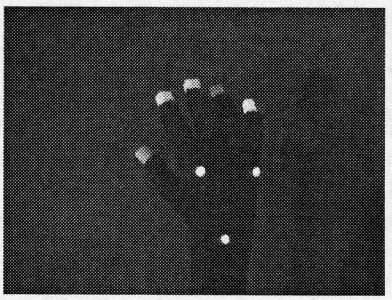


Figure 6: The experimental glove

A single camera records the hand's motion, while feature tracking is achieved by starting from the position of the cues in the previous frame. As mentioned above, color segmentation is based on the hue and saturation of the pixels under consideration.

3.2 Reconstruction of palm posture

Assuming the dimension of the palm's triangle is known, it is possible to compute the position of the 3 points in space according to a simple equation given in [10] under the orthoperspective camera projection approximation. The triangle's pose recovery problem generally leads to two distinct solutions which we discriminate with the estimated palm orientation.

3.3 Reconstruction of finger postures

3.3.1 Case of fingers II to V

Starting from an estimation of the abduction angle q_0 based on the hand's rest position, the finger's plane of motion is known, thus giving the position of the fingertip in space. The inverse kinematics equations then give joint values q_1, q_2, and q_3, as shown in Figure 7.

The abduction angle needs to be determined precisely in order not to lead to unreachable fingertip positions in space for the inverse kinematics problem. As a result, among the possible values for q_0, according to the hand's physiology, the corresponding flexion angle q_1

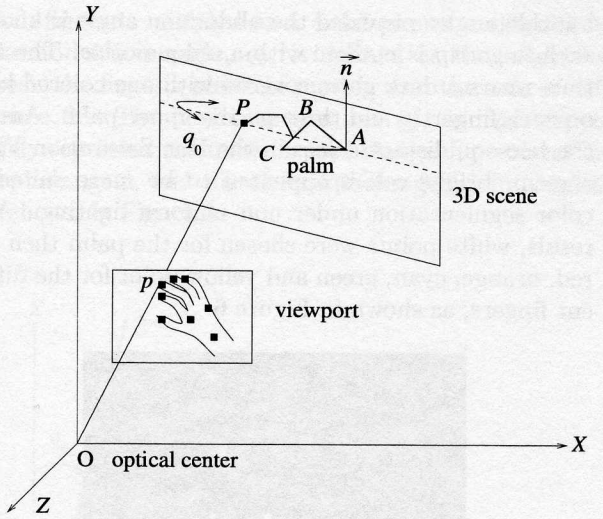


Figure 7: 3D reconstruction of the fingertips

is tested over the authorized scale. The mean angle is chosen so as to maximize the angular distance from the critical positions, Figure 8.

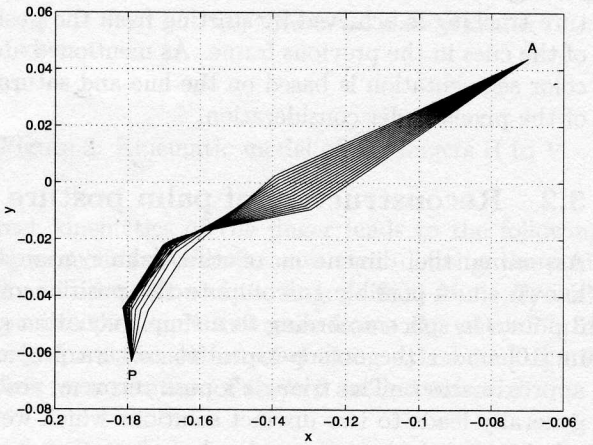


Figure 8: Multiple solutions to the inverse kinematics problem

3.3.2 Case of the thumb

The thumb's plane of motion is defined by q_0 and q_1. The intersection between this plane and the line of sight going through the thumb tip gives the estimated 3D position of the thumb. To ensure that this position matches the thumb's kinematics so that an inverse kinematics solution exists, we minimize the distance between the estimated 3D position of the thumb and the nearest position the thumb can reach with respect to its kinematics over q_0 and q_1. A smoothing is finally performed to give a realistic motion of the reconstructed hand when visualized using a graphical hand model.

3.4 Occlusion handling

During the cue detection phase, it may happen that one or more color tips are not detected. In this case, it is assumed that the visibility of the fingertip under consideration is obstructed by another finger. We define a *visibility table* which orders the 3D positions of the fingertips by their visibility to the camera. One fingertip may then be occluded by any fingertip that precedes it in the table. Observations of human movements reveal that three major classes of occlusions stand out. First, when the hand rotates around the middle finger axis, the furthest fingertips may be occluded by the fingers nearest to the camera, as shown in Figure 9(a,b). Second, the hand may rotate if the wrist is folding, resulting in self-occluded fingertips, as illustrated in Figure 9(c).

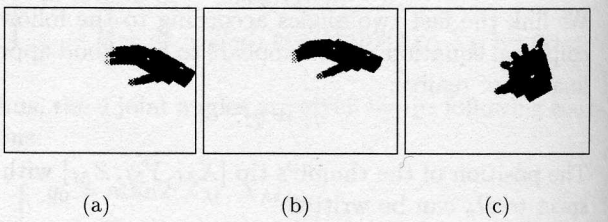


Figure 9: occlusions due to the palm motion

At last, when the fingers are closing, one finger may be occluded by any finger that precedes it in the visibility table. An example is shown in Figure 10.

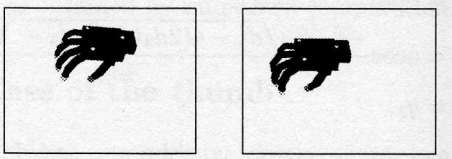


Figure 10: Occlusions due to the fist closure

During the reconstruction step, flexion angles of the occluded fingers are iteratively incremented until the occlusion is validated by testing the non visibility of the occluded tip. In order to validate the presence of any occlusions, geometric visibility tests must be performed on the 3D hand configuration. The fingertip is modeled as a sphere and each phalanx as a cylinder.

One fingertip may either be occluded by its own phalanxes or that of another finger. Figure 11 highlights the possibilities of occlusions of one sphere by another sphere (at the end of the cylinder) or by a cylinder (one phalanx), depending on the viewpoint.

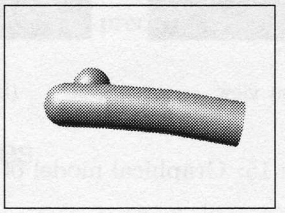


Figure 11: Phalanx and fingertip model

Self-occlusion: the angle between the line of sight reaching the fingertip and the motion plane of the finger must be less than the visibility angle of the fingertip with respect to the phalanxes.
Occlusion of one sphere by another sphere: the sphere must be inside the cone whose apex is the optic center and that wraps the first sphere.
Occlusion of one sphere by a cylinder: the sphere must be inside the dihedron defined by the two tangent planes of the cylinder going through the optic center.
The test of self-occlusion is performed first, followed by that of the tips and the phalanxes of all foremost fingers until the occlusion is validated. More details are covered in [11].

3.5 Illustration

In order to validate the hand reconstruction results, a graphical model of the hand simulates the motion thus obtained. The virtual hand is modeled with the graphic library OpenGL, taking advantage of the low-level object definition. The graphical hand model parameters are computed with the joint values extracted from the hand in the image sequence. The model duplicates the motion of the hand in the reference frame of the camera, allowing direct estimation of the reconstruction quality. Besides, the hand can be animated in any reference frame whose expression is given with respect to the reference frame of the camera, as shown in Figure 12 .

Despite some inaccuracy in the replica as far as the thumb is concerned, the initial posture of the hand is fairly well recovered by the graphical model. The difference in the thumb angles can be accounted for by the error in the 3D position of the thumb's tip owing to the optimization procedure of reconstruction. Besides, the main weakness of the system is its sensitivity to error in palm pose estimation. These errors are mostly due

(a) Original frame (b) Graphical model (c) Graphical model in a relative frame

Figure 12: Reconstruction of the hand in another reference frame

to the fact that the palm is in fact a non-rigid object and that deformations of the triangular pattern drawn on the glove are incorrectly interpreted by the pose estimation algorithm. Reducing the size of the pattern so that it should rest on a nearly-rigid part of the hand can greatly contribute to the elimination of this problem and is being investigated. Nevertheless, more errors occur in the reconstruction of a single static image than in an image extracted from a sequence where predictions of the current angles from the previous ones as well as smoothing enhance motion naturalness, as shown in Figure 13.

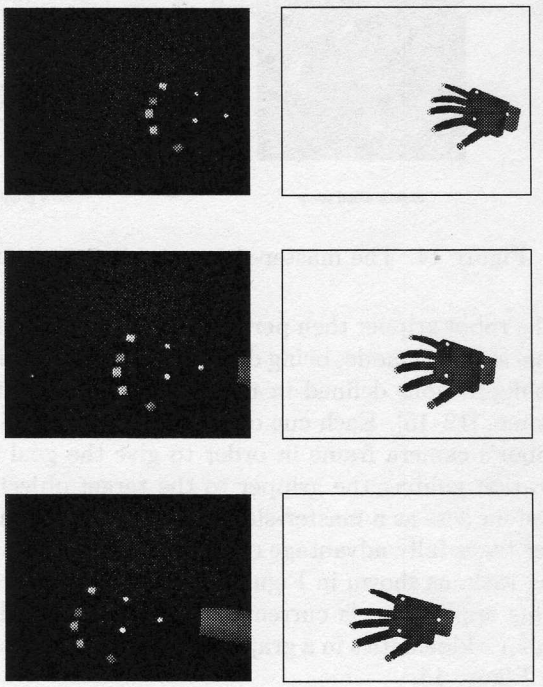


Figure 13: Reconstruction of an image sequence

As a result, it is possible to give a meaningful represen-

tation of the hand sequence from any point of view.

4 Applications

4.1 The hand as a visual control guide

The system we have just described may be exploited in several applications requiring the information extracted from the hand posture, in particular so called 'learning by watching tasks' involving manipulator control. We consider here a gripper control application where a camera is calibrated in the robot's reference frame. The image sequence of the hand can be computed in any reference frame defined with respect to the frame of the hand's camera. Therefore, the hand sequence can be projected in the frame of the robot's camera, provided the relationship between the hand's camera and the robot's camera is known.

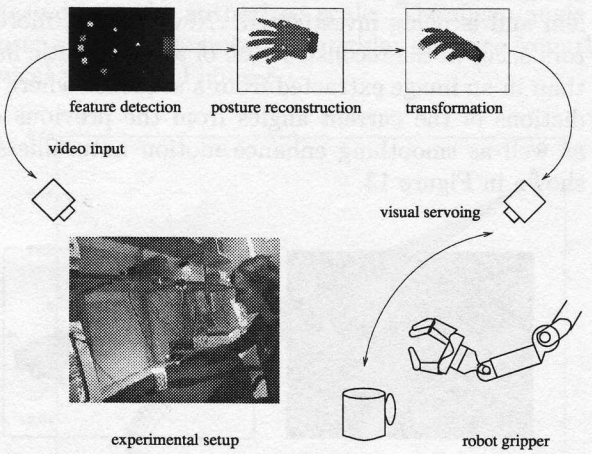
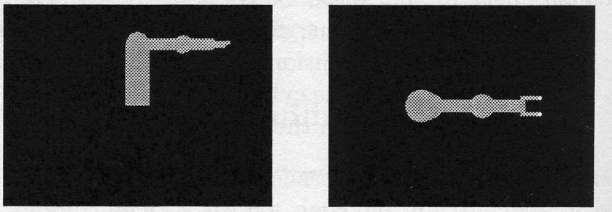


Figure 14: The master-slave servoing application

The robot gripper then performs a grasping task in a visual servoing mode, being driven by the passing through configurations defined in the robot's camera reference frame, [12–15]. Each cue of interest is projected in the robot's camera frame in order to give the goal configuration guiding the gripper to the target object. This system acts as a master-slave controler where the gripper takes fully advantage of the human skill in a grasping task, as shown in Figure 14.

This application is currently tested by simulating the robot's kinematics in a graphical environment, as shown in Figure 15.

For the moment, two white points are chosen on the top of the gripper to constitute the cues to be controlled in the image. The desired trajectory of the two reference points is simulated. The gripper is then controlled in visual servoing mode in order to track the desired cues,



(a) front view (b) top view

Figure 15: Graphical model of the robot

starting from a remote position. The resulting feature trajectory of the gripper is depicted in Figure 16, where co-ordinates are given in pixels. The real trajectory re-

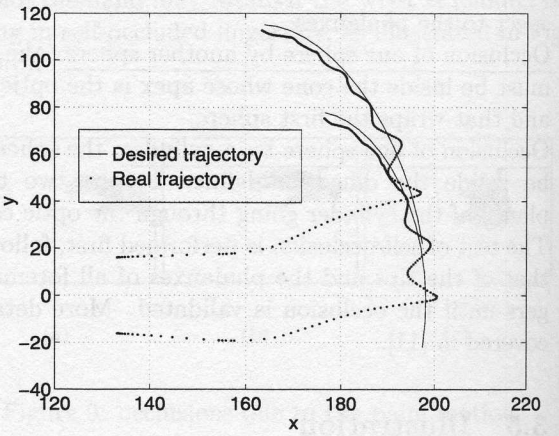


Figure 16: The trajectory of the gripper cues from visual servoing

veals some oscillations due to the lack of information given by two points as regards the gripper's orientation. Yet, more precision can be gained by investigating the control of at least three points on the gripper.

5 Conclusion

This article describes a hand posture reconstruction system taking in input the image sequence of a moving hand. The mathematical model of the hand is defined and the inverse kinematics solution is provided as well. The posture of the hand is recovered after robust detection and reconstruction is performed and a sequence of 3D graphic hands is generated. It becomes easy to display this sequence in any reference frame and in particular in the field of view of a robot gripper in order to achieve visual servoing control for the gripper. The

hand model is also accurate for 3D input device applications where a mapping is defined between the hand posture and some actions on a display, all the more since the amount of degrees of freedom of the hand allows no restriction. Further work is being carried out on the number and kind of 2D hand features for the visual servoing, so that more precise gripper control could be performed.

References

[1] V.I. Pavlovic, R. Sharma, and T.S. Huang, "Visual interpretation of hand gestures for Human-Computer Interaction: A review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 677–694, July 1997.

[2] J. Davis and M. Shah, "Recognizing hand gestures", in *Proceedings of the European Conference on Computer Vision*, Stockholm, Sweden, 1994, pp. 331–340.

[3] T. Darrell and A. Pentland, "Space-time gestures", in *IEEE conference on Computer Vision and Pattern Recognition*, New York, 1993, pp. 335–340.

[4] L. Goncalves, E. Di Bernardo, E. Ursella, and P. Perona, "Monocular tracking of the human arm in 3D", in *Proceedings of the IEEE International Conference on Computer Vision*, Cambridge, MA, 1995, pp. 764–770.

[5] J.J. Kuch and T.S. Huang, "Vision based hand modeling and tracking for virtual teleconferencing and telecollaboration", in *Proceedings of the IEEE International Conference on Computer Vision*, Cambridge, MA, 1995, pp. 666–671.

[6] J.M. Rehg and T. Kanade, "DigitEyes: Vision-based human hand tracking", Tech. Rep. CMU-CS-93-220, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, December 1993.

[7] J. Lee and T. Kunii, "Model-Based Analysis of Hand Posture", *IEEE Computer Graphics and Applications*, vol. 15, no. 5, pp. 77–86, 1995.

[8] R. Tubiana, *The Hand*, vol. 1, Sanders, Philadelphia, PA, 1981.

[9] H. Rijpkema and M. Girard, "Computer animation of knowledge-based human grasping", *Computer Graphics, Proc. SIGGRAPH*, vol. 25, no. 4, pp. 339–347, 1991.

[10] D. DeMenthon and L.S. Davis, "New exact and approximate solutions of the three-point perspective problem", in *Proceedings of the IEEE International Conference on Robotics and Automation*, Cincinnati, OH, 1990, vol. 1, pp. 40–45.

[11] Fabienne Lathuilière, "Visual tracking of hand posture with occlusion handling", Tech. Rep. GRPR-RT 9902, GRPR, École Polytechnique de Montréal, March 1999.

[12] J.T. Feddema, C.S.G. Lee, and O.R. Mitchell, "Weighted selection of image features for resolved rate visual feedback control", *IEEE Transactions on Robotics and Automation*, vol. 7, no. 1, pp. 31–47, 1991.

[13] J.T. Feddema and O.R. Mitchell, "Vision-guided servoing with feature-based trajectory generation", *IEEE Transactions on Robotics and Automation*, vol. 5, no. 5, pp. 691–700, 1989.

[14] L.E. Weiss, A.C. Sanderson, and C.P. Neuman, "Dynamic sensor-based control of robots with visual feedback", *IEEE Journal of Robotics and Automation*, vol. 3, no. 5, pp. 404–417, 1987.

[15] B. Espiau, F. Chaumette, and P. Rives, "A new approach to visual servoing in robotics", *IEEE Transactions on Robotics and Automation*, vol. 8, no. 3, pp. 313–326, 1992.